BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 0 649 144 B1

(12) EUROPEAN PATENT SPECIFICATION

(45) Date of publication and mention of the grant of the patent:
05.01.2000 Bulletin 2000/01

(51) Int Cl.[7]: **G11B 27/28**, G11B 27/10, G11B 27/028

(21) Application number: **94114025.3**

(22) Date of filing: **07.09.1994**

(54) Automatic indexing of audio using speech recognition

Automatische Indizierung von Audiosignalen mittels Spracherkennung

Indexation automatique de signaux audio au moyen de reconnaissance de la parole

(84) Designated Contracting States:
DE FR GB

(30) Priority: **18.10.1993 US 138949**

(43) Date of publication of application:
19.04.1995 Bulletin 1995/16

(73) Proprietor: **International Business Machines Corporation**
Armonk, N.Y. 10504 (US)

(72) Inventors:
- **Ellozy, Hamed A.**
  Bedford Hills, New York 10507 (US)
- **Kanevsky, Dimitri**
  Ossining, New York 10568 (US)
- **Kim, Michelle Y.**
  Scarsdale, New York 10583 (US)
- **Nahamoo, David**
  White Plains, New York 10605 (US)
- **Picheny, Michael A.**
  White Plains, New York 10606 (US)
- **Zadrozny, Wlodek W.**
  Mohegan Lake, New York 10547 (US)

(74) Representative: **Teufel, Fritz, Dipl.-Phys.**
IBM Deutschland Informationssysteme GmbH,
Patentwesen und Urheberrecht
70548 Stuttgart (DE)

(56) References cited:
EP-A- 0 507 743
US-A- 5 136 655
US-A- 5 199 077
WO-A-92/11634
US-A- 5 149 104

- **IBM TECHNICAL DISCLOSURE BULLETIN., vol.33, no.10A, March 1991, NEW YORK US pages 295 - 296, XP110048 'CORRELATING AUDIO AND MOVING-IMAGE TRACKS'**

EP 0 649 144 B1

Printed by Jouve, 75001 PARIS (FR)

similarity is trained and used for Viterbi alignment.)

c) Use the alignment result to break the summary into segments each associated with an anchor point. Since the anchor points carry time stamps, we achieve a time alignment between the summary script and the speech data.

d) Repeat this process on the subsegments that can still be broken into smaller parts.

[0072] The following is an explanation of Figure 5. The block 401 contains a decoded text (ordered series of recognized words) DT that is schematically represented by a vertical left series of words 1,2,3, ...8 and a transcript T that is schematically represented by a vertical right series of words 1',2',3' ... 7'. The pairs of words (1,1'), (4,5'), (8,7') were matched as described in Figure 4. The series of words 1,2,....8 is aligned against audio data (block 402) in the course of decoding (Figure 3 block 42), as schematically shown inside block 402. Let (T0, T1), (T1, T2),...(T7, T8) correspond to the beginnings and ends of words 1,2,3...8, respectively. Then the matched transcript words 1', 5', 7' will correspond to time data (T0,T1), (T3,T4), (T7,T8), respectively (via the matched decoded words).

[0073] Remaining decoded words can be aligned with the time data by linear interpolation. For example, time segment (T1, T3) corresponds to the word segment W2, W3, and can be aligned in accordance with the length of words. For example, if W2 consists of N phonemes and W3 of M phonemes, and t-T3-TI then the segment ýTI, T1+t*N/(N+M) corresponds to W2, and the segment ýT1+*N/(N+M), T3* corresponds to W3.

[0074] The aligned transcript-audio data is transferred to the block 403 where is it aligned with video data from the record/playback deck 19 of Figure 3. This alignment is obtained by time stamping that was done for audio-video data.

[0075] The following is an explanation of Figure 6 in which the speech recognizer vocabulary is obtained from segments of the text transcript. The block 501 contains the current part of a transcript Ti that is processed. This part of the transcript Ti is used to derive the vocabulary V 504 from which the text in Ti was formed, and the approximate size 503 of the tape section 505 that contains the speech that corresponds to Ti. The size can be obtained estimating for each word Wr in Ti the maximum possible size Dr of its corresponding audio data on the tape, and taking the sum of all Dr (r=1,2...) in a segment as the length of a segment in the tape.

[0076] This information is transferred to the block 502 where the following tasks are performed. The end of the audio segment on the tape that corresponds to the previous T(i-1) text (or the beginning of the tape for the first T1 segment) is identified. the next segment of the tape with length that is provided from the block 501 is played automatic speech recognizer 506. The automatic speech recognizer decodes this audio data using the vocabulary that was provided from the block 501. The automatic speech recognizer sends each decoded series of words W1,W2, ... Wk to the block 501 and compares it with the text Ti. If the decoded series of words matches well with the corresponding part V1, V2,...V1 in Ti, then the next word V(1+1) is added to the list of alternative words the automatic speech recognizer is processing in decoding the corresponding segment of audio data. (This candidate word V(1+1) could be given with an additional score that represents the likelihood of being the next word in the considered path). After the whole text Ti is decoded, the end of the tape audio data that corresponds to the end of the text is defined. This end of the audio segment is transferred to the next step (decoding of T(i+1)) part of the text if Ti is not the last segment in T.

Claims

1. An apparatus for indexing an audio recording comprising:

   an acoustic recorder (70) for storing an ordered series of acoustic information signal units representing sounds generated from spoken words, said acoustic recorder having a plurality of recording locations, each recording location storing at least one acoustic information signal unit ;

   a speech recognizer (72) for generating an ordered series of recognized words having a high conditional probability of occurrence given the occurrence of the sounds represented by the acoustic information signals, each recognized word corresponding to at least one acoustic information signal unit, each recognized word having a context of at least one preceding or following recognized word;

   a text storage device (74) for storing an ordered series of index words, said ordered series of index words comprising a visual representation of at least some of the spoken words represented by the acoustic information signal units, each index word having a context of at least one preceding or following index word; and

   means (76) for comparing the ordered series of recognized words with the ordered series of index words to pair recognized words and index words which are the same word and which have matching contexts, and for tagging each paired index word with the recording location of the acoustic information signal unit corresponding to the recognized word paired with the index

word.

2. An apparatus as claimed in Claim 1, characterized in that the speech recognizer aligns each recognized word with at least one acoustic information signal unit.

3. An apparatus as claimed in Claim 1 or 2, characterized in that:

each recognized word which is not paired with an index word has a nearest preceding paired recognized word in the ordered series of recognized words, and has a nearest following paired recognized word in the ordered series of recognized words;

the context of a target recognized word comprises the number of other recognized words preceding the target recognized word and following the nearest preceding paired recognized word in the ordered series of recognized words;

the context of a target index word comprises the number of other index words preceding the target index word and following the nearest preceding paired index word in in the ordered series of index words; and

the context of a recognized word matches the context of an index word if the context of the recognized word is within a selected threshold value of the context of the index word.

4. A method of indexing an audio recording comprising:

storing an ordered series of acoustic information signal units representing sounds generated from spoken words, said acoustic recorder having a plurality of recording locations, each recording location storing at least one acoustic information signal unit;

generating an ordered series of recognized words having a high conditional probability of occurrence given the occurrence of the sounds represented by the acoustic information signals, each recognized word corresponding to at least one acoustic information signal unit, each recognized word having a context of at least one preceding or following recognized word;

storing an ordered series of index words, said ordered series of index words comprising a visual representation of at least some of the spoken words represented by the acoustic information signal units, each index word having a context of at least one preceding or following index word;

comparing the ordered series of recognized words with the ordered series of index words to pair recognized words and index words which are the same word and which have matching contexts; and

tagging each paired index word with the recording location of the acoustic information signal unit corresponding to the recognized word paired with the index word.

5. A method as claimed in claim 4, characterized in that:

each recognized word comprises a series of one or more characters;

each index word comprises a series of one or more characters; and

a recognized word is the same as an index word when both words comprise the same series of characters.

6. A method as claimed in any one of the preceding claims, characterized in that:

the context of a target recognized word comprises the number of other recognized words preceding the target recognized word in the ordered series of recognized words;

the context of a target index word comprises the number of other index words preceding the target index word in the ordered series of index words; and

the context of a recognized word matches the context of an index word if the context of the recognized word is within a selected threshold value of the context of the index word.

7. A method as claimed in any one of the preceding claims, characterized in that:

each index word which is not paired with a recognized word has a nearest preceding paired index word in the ordered series of index words, and has a nearest following paired index word in the ordered series of index words; and

the step of tagging comprises tagging a non-paired index word with a recording location be-

tween the recording location of the nearest preceding paired index word and the recording location of the nearest following paired index word.

8. A method as claimed in any one of the preceding claims, further comprising the step of aligning each recognized word with at least one acoustic information signal unit.

9. A method as claimed in any of the preceding claims, characterized in that:

each recognized word which is not paired with an index word has a nearest preceding paired recognized word in the ordered series of recognized words, and has a nearest following paired recognized word in the ordered series of recognized words;

the context of a target recognized word comprises the number of other recognized words preceding the target recognized word and following the nearest preceding paired recognized word in the ordered series of recognized words;

the context of a target index word comprises the number of other index words preceding the target index word and following the nearest preceding paired index word in in the ordered series of index words; and

the context of a recognized word matches the context of an index word if the context of the recognized word is within a selected threshold value of the context of the index word.

Patentansprüche

1. Eine Vorrichtung, um eine Audioaufzeichnung zu indizieren, mit

einem Tonaufzeichnungsgerät (70), um eine Reihenfolge von akustischen Informationssignaleinheiten zu speichern, die von gesprochenen Worten erzeugte Töne darstellen, wobei das Tonaufzeichnungsgerät eine Vielzahl von Aufzeichnungsstellen hat, von denen jede Aufzeichnungsstelle mindestens eine akustische Informationssignaleinheit speichert;

einer Spracherkennung (72), um eine Reihenfolge von erkannten Worten mit einer hohen, bedingten Wahrscheinlichkeit des Auftretens zu erzeugen, die von dem Auftreten der Töne gegeben wird, die von den akustischen Informationssignalen dargestellt werden, wobei jedes erkannte Wort mindestens einer akustischen Informationssignaleinheit entspricht, und jedes erkannte Wort einen Kontext zu mindestens einem vorhergehenden oder nachfolgenden erkannten Wort hat;

einem Textspeichergerät (74), um eine Reihenfolge von Indexwörtern zu speichern, wobei die Reihenfolgen mit Indexwörtern eine visuelle Darstellung von mindestens einem der gesprochenen Worte enthalten, die von den akustischen Informationssignaleinheiten dargestellt werden, und jedes Indexwort einen Kontext zu mindestens einem vorhergehenden oder nachfolgenden Indexwort hat; und

Mittel (76), um die Reihenfolgen von erkannten Wörtern mit den Reihenfolgen von Indexwörtern zu vergleichen, um die erkannten Wörter und Indexwörter, die gleich sind und übereinstimmende Kontexte haben, zu paaren, und um jedes gepaarte Indexwort in der Aufzeichnungsstelle der akustischen Informationssignaleinheit entsprechend dem erkannten Wort, das mit dem Indexwort gepaart wurde, zu kennzeichnen.

2. Eine Vorrichtung wie in Anspruch 1 angemeldet, dadurch gekennzeichnet, daß die Spracherkennung jedes erkannte Wort zu wenigstens einer akustischen Informationssignaleinheit ausrichtet.

3. Eine Vorrichtung wie in Anspruch 1 oder 2 angemeldet, dadurch gekennzeichnet, daß

jedes erkannte Wort, das nicht mit einem Indexwort gepaart ist, ein am nächsten kommendes Wort hat, das vor dem gepaarten, erkannten Wort in den Reihenfolgen der erkannten Wörter liegt, und ein am nächsten kommendes Wort hat, das nach dem gepaarten, erkannten Wort in den Reihenfolgen der erkannten Wörter liegt;

der Kontext eines erkannten Zielworts die Anzahl der anderen erkannten Wörter enthält, die dem erkannten Zielwort vorausgehen und die dem am nächsten kommenden Wort, das vor dem gepaarten, erkannten Wort in den Reihenfolgen der erkannten Wörter liegt, folgen;

der Kontext eines Zielindexworts die Anzahl der anderen Indexwörter enthält, die dem Zielindexwort vorausgehen, und die dem am nächsten kommenden Wort, das vor dem gepaarten Indexwort in den Reihenfolgen der Indexwörter liegt, folgen; und

Wort in den Reihenfolgen der erkannten Wörter liegt, und ein am nächsten kommendes Wort hat, das nach dem gepaarten, erkannten Wort in den Reihenfolgen der erkannten Wörter liegt;

der Kontext eines erkannten Zielworts die Anzahl der anderen erkannten Wörter enthält, die dem erkannten Zielwort vorausgehen und die dem am nächsten kommenden Wort, das vor dem gepaarten, erkannten Wort in den Reihenfolgen der erkannten Wörter liegt, folgen;

der Kontext eines Zielindexworts die Anzahl der anderen Indexwörter enthält, die dem Zielindexwort vorausgehen, und die dem am nächsten kommenden Wort, das vor dem gepaarten Indexwort in den Reihenfolgen der Indexwörter liegt, folgen; und

der Kontext eines erkannten Worts mit dem Kontext eines Indexworts übereinstimmt, wenn der Kontext des erkannten Worts innerhalb eines ausgewählten Schwellwerts des Kontexts vom Indexwort liegt.

Revendications

1. Un dispositif pour indexer un enregistrement audio, comprenant :

un enregistreur acoustique (70), destiné à stocker une série classée d'unités de signal d'information acoustique représentant des sons générés à partir de mots énoncés, ledit enregistreur acoustique ayant une pluralité d'emplacements d'enregistrement, chaque emplacement d'enregistrement stockant au moins une unité de signal d'information acoustique;

un identificateur de la parole (72) destiné à générer une série classée de mots identifiés ayant une forte probabilité conditionnelle d'occurrence, étant donné l'occurrence des sons représentés par les signaux d'information acoustique, chaque mot identifié correspondant à au moins une unité de signal d'information acoustique, chaque mot identifié ayant un contexte d'au moins un mot identifié précédant ou suivant;

un dispositif de stockage de texte (47) destiné à stocker une série classée de mots d'indexation, lesdites séries classées de mots d'indexation comprenant une représentation visuelle d'au moins certains des mots énoncés, représentés par les unités de signal d'information acoustique, chaque mot d'indexation ayant un contexte d'au moins un mot d'indexation précédant ou suivant; et

des moyens (76) pour comparer les séries classées de mots identifiés avec les séries classées de mots d'indexation, à des paires de mots identifiés et de mots d'indexation qui sont le même mot et qui ont des contextes de coïncidence, et pour étiqueter chaque mot d'indexation en paires avec l'emplacement d'enregistrement de l'unité de signal d'information acoustique correspondant au mot identifié mis en paire avec le mot d'indexation.

2. Un dispositif selon la revendication 1, caractérisé en ce que l'identificateur de la parole aligne chaque mot identifié avec au moins une unité de signal d'information acoustique.

3. Un dispositif selon la revendication 1 ou 2, caractérisé en ce que :

chaque mot identifié qui n'est pas mis en paire avec un mot d'indexation a un mot identifié mis en paire précédent le plus proche dans les séries classées de mots identifiés, et a un mot identifié mis en paire suivant le plus proche dans les séries classées de mots identifiés;

le contexte d'un mot identifié cible comprend le nombre d'autres mots identifiés qui précèdent le mot identifié cible et qui suivent le mot identifié mis en paire précédent le plus proche dans les séries classées de mots identifiés;

le contexte d'un mot d'indexation cible comprend le nombre des autres mots d'indexation qui précèdent le mot d'indexation cible et qui suivent le mot d'indexation mis par paire précédent le plus proche dans les séries classées de mots d'indexation; et

le contexte d'un mot identifié coïncide avec le contexte d'un mot d'indexation si le contexte dans le mot identifié est situé dans une valeur de seuil sélectionné du contexte du mot d'indexation.

4. Un procédé d'indexation d'un enregistrement audio comprenant :

la mémorisation d'une série classée d'unités de signal d'information acoustique représentant des sons générés par des mots énoncés, ledit enregistreur acoustique ayant une pluralité d'emplacements d'enregistrement, chaque emplacement d'enregistrement stockant au moins une unité de signal d'information acous-

tique;

la génération de séries classées de mots identifiés ayant une forte probabilité conditionnelle d'occurrence étant donné l'occurrence des sons représentés par les signaux d'information acoustique, chaque mot identifiés correspondant à au moins une unité de signal d'information acoustique, chaque mot identifié ayant un contexte d'au moins un mot identifié précédant ou suivant;

la mémorisation de séries classées de mots d'indexation, ladite série classée de mots d'indexation comprenant une représentation visuelle d'au moins certains des mots énoncés, représentés par les unités de signal d'information acoustique, chaque mot d'indexation ayant un contexte d'au moins un mot d'indexation précédant ou suivant;

la comparaison des séries classées des mots identifiés avec les séries classées des mots d'indexation à des mots identifiés et des mots d'indexation mis en paires, qui sont le même mot et qui ont des contextes de coïncidence; et

l'étiquetage de chaque mot d'indexation mis en paire avec l'emplacement d'enregistrement de l'unité de signal d'information acoustique correspondant au mot identifié mis en paire avec le mot d'indexation.

5. Un procédé selon la revendication 4, caractérisé en ce que :

chaque mot identifié comprend une série d'un ou plusieurs caractères;

chaque mot d'indexation comprend une série d'un ou plusieurs caractères; et

un mot identifié est le même qu'un mot d'indexation lorsque les deux mots sont constitués des mêmes séries de caractères.

6. Un procédé selon l'une quelconque des revendications précédentes, caractérisé en ce que :

le contexte d'un mot identifié cible comprend le nombre d'autres mots identifiés précédant le mot identifié cible dans les séries classées des mots identifiés;

le contexte d'un mot d'indexation cible comprend le nombre d'autres mots d'indexation précédant le mot d'indexation cible dans les séries classées des mots d'indexation; et

le contexte d'un mot identifié coïncidant avec le contexte d'un mot d'indexation si le contexte du mot identifié est situé dans une valeur seuil sélectionnée du contexte du mot d'indexation.

7. Un procédé selon l'une quelconque des revendications précédentes, caractérisé en ce que :

chaque mot d'indexation qui n'est pas mis en paire avec un mot identifié a un mot d'indexation mis en paire précédant le plus proche dans les séries classées des mots d'indexation, et un mot d'indexation mis en paire suivant le plus proche dans les séries classées des mots d'indexation; et

l'étape d'étiquetage comprend l'étiquetage d'un mot d'indexation non-mis en paire avec un emplacement d'enregistrement entre l'emplacement d'enregistrement du mot d'indexation mis en paire précédant le plus proche et l'emplacement d'enregistrement du mot d'indexation mis en paire suivant le plus proche.

8. Un procédé selon l'une quelconque des revendications précédentes, comprenant en outre l'étape consistant à aligner chaque mot identifié avec au moins une unité de signal d'information acoustique.

9. Un procédé selon l'une quelconque des revendications précédentes, caractérisé en ce que :

chaque mot identifié qui n'est pas mis en paire avec un mot d'indexation a un mot identifié mis en paire précédant le plus proche dans les séries classées de mots identifiés, et à un mot identifié mis en paire suivant le plus proche dans les séries classées de mots identifiés;

le contexte d'un mot identifié cible comprend le nombre d'autres mots identifiés précédant le mot identifié cible et suivant le mot identifié mis en paire précédant le plus proche dans les séries classées de mots identifiés;

le contexte d'un mot d'indexation cible comprend le nombre d'autres mots d'indexation précédant le mot d'indexation cible et suivant le mot d'indexation mis en paire précédent le proche dans les séries classées de mots d'indexation; et

le contexte d'un mot identifié coïncide avec le contexte d'un mot d'indexation si le contexte du mot identifié est situé dans une valeur seuil sélectionnée du contexte du mot d'indexation.